

专利信息本体的设计及应用

缪涵琴, 孙涌

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要:讨论了本体与国际专利分类表的内在联系。在专利信息服务系统中,以国际专利分类表为基础,提出了符合专利信息特点的IPC本体;设计了采用本体检索与关键字检索有效结合的组合检索方式;增加了从企业用户信息中提取出的用户兴趣模型。把用户兴趣模型转化为符合系统检索规则的检索条件加入到原组合检索中,进一步缩小检索范围,在一定程度上提高了专利信息检索的查全率和查准率。

关键词:国际专利分类表;专利信息服务系统;本体;组合检索;用户兴趣模型

中图分类号:TP301.2

文献标识码:A

文章编号:1673-629X(2007)12-0204-03

The Application of Ontology in Patent Information Services System

MIAO Han-qin, SUN Yong

(Computer Science & Technology Institute, Soochow University, Suzhou 215006, China)

Abstract: Discusses the inherent relation between ontology and international patent classification. On the basis of international patent classification, the Patent Information Services System puts forward a IPC ontology which addresses the characteristics of patent information, and combines the ontology with keywords retrieval, and adds user's interest model which extracts from corporation information. This model translates to condition which is true of the retrieval regular of system and joins the combined search mode formerly. This reduces the bound of searching and improves the recall and the precision of patent information retrieval.

Key words: international patent classification; patent information services system; ontology; combined retrieval; user's interest model

0 引言

专利是一种财产权,是运用法律保护手段独占市场、抢占潜在市场的有力武器。专利信息检索则是根据一项数据特征,从大量的专利数据库中挑选符合某一特定要求的文献或信息的过程。专利信息检索是一项复杂的工作,由多种因素构成,如:检索方式,检索目的,检索经验等。这些因素共同制约着专利信息检索的过程,直接影响着专利信息检索的结果。企业可以通过专利信息检索,了解行业的最新技术动态以及发展趋势,为企业技术人员确定研发方向和管理层选择投资项目提供有价值的参考,为企业的快速发展提供帮助。

目前所使用的专利信息检索服务主要集中在关键字匹配方式^[1,2]。基于关键字的检索容易造成信息冗余和信息丢失等问题。专利信息的特点是数据量大、

类别多,所以对检索提出了更高的要求。本体作为一种新的知识表示方式,将知识表示扩展到语义的层次上。目前国内外比较有影响的专利信息检索系统中,没有涉及到本体的使用。笔者将讨论如何将本体与专利信息的特点结合起来,提高专利信息检索的准确性,为企业提供有效的专利信息服务,包括帮助企业掌握最新技术情况,节省技术研发时间,为企业的长期发展提供科学的决策依据。

1 本体与国际专利分类法

自1998年现代因特网之父 Tim Berners-Lee 提出语义网(Semantic Web)概念起,语义网的研究逐渐受到人们的广泛关注。本体(Ontology)作为语义网的基石,在语义网的研究中具有非常重要的作用。

本体的概念最初起源于哲学领域,由亚里士多德提出,原意是指关于存在及其本质和规律的学说,后来随着人工智能的发展,被给予了新的定义,特指对概念化的一个显式的规格说明。最为著名的是1993年 Gruber 提出的本体是概念化的明确的规范说明^[3]。其中概念化(conceptualization)是现实世界中现象的抽象

收稿日期:2007-02-01

基金项目:2005年度教育部科研重点项目(205059)

作者简介:缪涵琴(1981-),女,江苏苏州人,硕士研究生,研究方向为智能信息处理、数据库;孙涌,副教授,博士,研究方向为软件工程、数据库。

模型,要明确标识与现象相关的概念;明确(explicit)的意思是指被使用概念的类型以及概念在使用中的约束被明确地定义出来;形式化(formal)的意思是指本体应该是机器可读的;共享(share)是反映本体中的知识是中立的、一致认可的^[4,5]。

本体能够将领域中的各种概念及相关关系显式地、形式化地表达出来,从而将术语的语义显式地表达出来,因而在语义查询方面发挥着重要的作用。

从 1968 年 9 月 1 日起世界专利国家开始使用“国际专利分类法”统一专利文献分类,为专利文献的检索提供了极大的方便^[6]。世界上主要的专利分类体系可归纳为以下四种:国际专利分类表、美国专利分类表、英国专利分类表、英国德温特出版公司编制的分类表。

目前国际上主要采用的分类表是国际专利分类表,因为它兼顾了专利审查和公众查找专利情报的要求,采用应用分类和功能分类相结合的混合分类结构^[7]。

国际专利分类法的作用是^[8]:

(1)使各国专利文献获得统一的分类工具,以便于对专利文献进行分类、管理、使用、查找;

(2)用于在各种检索中使用,如为确定专利申请的新颖性、创造性(包括对技术先进性和实用价值作出评价)而进行专利文献检索时的一种有效检索工具;

(3)利用 IPC 分类表编排专利文献,使用者可方便地从中获得技术上和法律上的信息;

(4)作为对所有专利信息使用者进行选择性的基础;

(5)作为对某一个技术领域进行现有技术水平调研的基础;

(6)作为进行专利统计工作的基础,从而对各个技术领域的技术发展状况作出评价。

文中将利用专利信息比较完善的分类体系,通过分析国际专利分类体系的分类规则,提出了把分类体系本体化的方法,并将其应用于专利信息的检索,通过扩充检索条件,得到更多相关或相近的检索结果。

2 引入本体后的系统设计

2.1 本体设计

国际专利分类(International Patent Classification, IPC)是世界各国专利机构都采用的专利分类方法,有统一的组织形式。IPC 分类体系是由高到低依次排列的等级式结构,设置的顺序是:部、大类、小类、主组、分组。其中每一个部的类号用一个大写英文字母标示,如 A 部, B 部。每一个大类的类号由部的类号及在其后加上两位数字组成,如: A01, A61。每一个小类类号由大类类号加上一个大写字母组成,如: A21B。由小类类号加上一个一位到三位的数、斜线“/”及数字“00”组成,如: A01B3/00, B62K19/00, C07C211/00。主组可以细分成若干小组。每一个小组的类号由小类类号加上一个一位至三位数,后面跟着斜线“/”符号,再加上一个除“00”以外的至少有两位的数组成,如: A23L1/32。根据 IPC 分类号的形式,本系统中所设计的本体结构如图 1 所示。

图中的第一层为 IPC 分类号中的部,第二层为大类,第三层为小类,第四层为主组,第五层为大分组,第六层为小组。

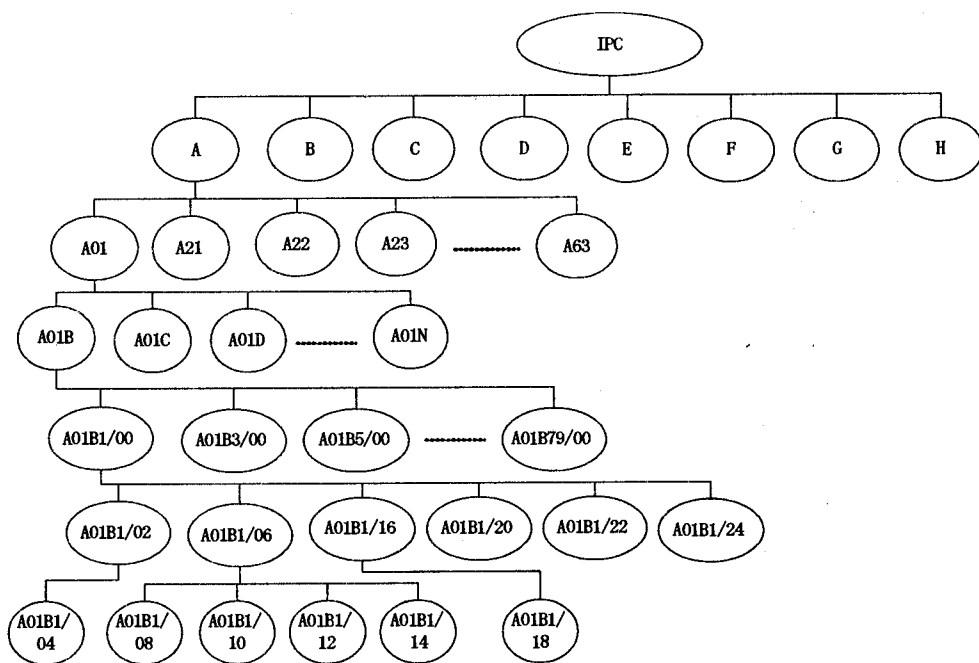


图 1 IPC 分类号的本体结构

根据用户输入的 IPC 分类号的不完整度,查找不同的层次,可以找到某些节点满足用户的输入条件,也可以得到这些节点的兄弟结点、父结点、孩子结点等,从而实现扩展。例如用户输入的是 A01B1/10,查 IPC 分类表可知为“带双铲刀或多铲刀的锄;手动中耕器”,从本体中可以得到该结点的兄弟结点 A01B1/08“带单铲刀的锄;手动中耕器”、A01B1/12“带齿铲刀的锄;手

动中耕器”、A01B1/14“只带齿的锄;手动中耕器”。

2.2 系统体系结构

引入本体后专利信息检索体系结构如图 2 所示。

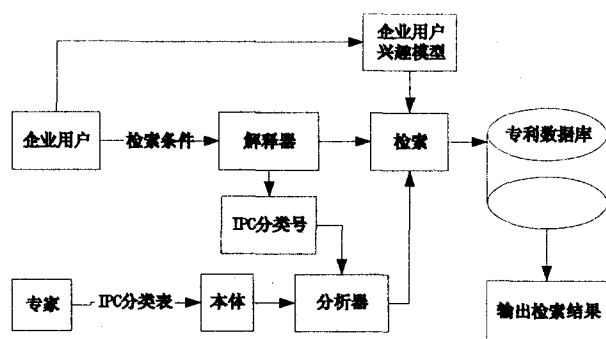


图 2 系统体系结构图

实现加入本体后的专利信息检索包括如下步骤。

(1) 专家根据 IPC 分类号创建本体。

在上一小节中已经介绍了根据 IPC 分类号的等级式结构来创建本体的过程,因为 IPC 分类号对于专利信息而言,是一种十分完善的分类体系,而且国际通用。这不仅保证了本体的普遍适用性和实用性,也遵守了专利信息的特殊性。

(2) 通过企业用户的信息建立用户兴趣模型。

每一个企业都有其特定的一个或若干个研究方向,可以从企业注册信息和与企业的沟通中获取企业的兴趣点,建立用户兴趣模型。例如某企业是生产电梯的,那么可以定义该企业的兴趣模型为:电梯/ab。

(3) 企业用户输入检索表达式。

企业用户想获取某些专利信息,可以通过将限制条件转化为规范的检索表达式来查询。例如,用户输入的检索表达式为:(王 and 李??)/in and (-20050505 and 20030505 -)/ad and A01B1/06/ic1。

(4) 从检索表达式中分离出 IPC 分类号。

从上面的检索表达式中分离出 IPC 分类号为 A01B1/06,由于专利信息的特殊性,专利文摘由各个字段组成,其中 ic1 字段中保存的就是 IPC 分类号。

(5) 通过本体将分离出的 IPC 分类号进行扩展。

根据本体可以得到 A01B1/06(“锄;手动中耕器”)结点的孩子结点 A01B1/08(“带单铲刀的”)、A01B1/10(带双铲刀或多铲刀的”)、A01B1/12(“带齿铲刀的”)、A01B1/14(“只带齿的”)等。A01B1/06 结点的兄弟结点为 A01B1/02(“锹;铲”)、A01B1/16(“拔草工具”)、A01B1/20(“不同类型手动工具的组合”)、A01B1/22(“把铲刀或类似工作部件固接在手柄上;可更换或可调整的铲刀”)、A01B1/24(“处理草地或草坪用的”)等。

主要代码如下:

... ..

```

OntModel m = ModelFactory.createOntologyModel();
OntDocumentManager dm = m.getDocumentManager();
dm.addAltEntry("http://www.owl-ontologies.com/ipc.owl",
"file:d:/ipcOnto/IPC.owl");
m.read("http://www.owl-ontologies.com/ipc.owl");
OntClass camera = m.getOntClass(camNS + name);
... ..
for (Iterator i = camera.listSubClasses(); i.hasNext(); ) {
    OntClass c = (OntClass) i.next();
    list.add(c.getLocalName());
}
... ..

```

(6) 扩展的 IPC 结合用户输入的检索表达式和用户兴趣模型检索数据库。

将用户输入的检索表达式与 IPC 分类号的扩展相结合,即表达式转化为:(王 and 李??)/in and (-20050505 and 20030505 -)/ad and (A01B1/06 or A01B1/08 or A01B1/10 or A01B1/12 or A01B1/14 or A01B1/02 or A01B1/16 or A01B1/20 or A01B1/22 or A01B1/24) /ic1。再考虑用户兴趣模型则将表达式转化为:(王 and 李??)/in and (-20050505 and 20030505 -)/ad and (A01B1/06 or A01B1/08 or A01B1/10 or A01B1/12 or A01B1/14 or A01B1/02 or A01B1/16 or A01B1/20 or A01B1/22 or A01B1/24) /ic1 and 电梯/ab and 日本/pa。最后将该表达式转化为标准的 SQL 语句,查询数据库。

3 结 语

文中在考虑专利信息的特殊性的前提下,提出了将国际专利分类号本体化,设计了专利信息所特有的 IPC 本体。今后的一项重要工作就是专利信息本体的进化,从理论上说,本体进化是必需的,也是可行的,但在本体的实际建设过程中看,本体的进化是一个比较复杂、难度较大的问题,是包含了众多方面相互作用和相互影响的复杂过程^[9]。

为了更好地达到检索的目的,利用关键字检索和本体检索相结合的方法,实现 IPC 分类号的语义扩展。IPC 分类号的扩展在一定程度上提高了检索的查全率,采用本体检索与关键字检索相结合的组合检索方式,并且考虑了通过企业信息建立的用户兴趣模型,又在一定程度上提高了检索的查准率。

参考文献:

[1] 周 园. 因特网上中国专利检索系统的比较分析[J]. 西安

(下转第 209 页)

(5)中断处理子过程:过程如图 2 所示。

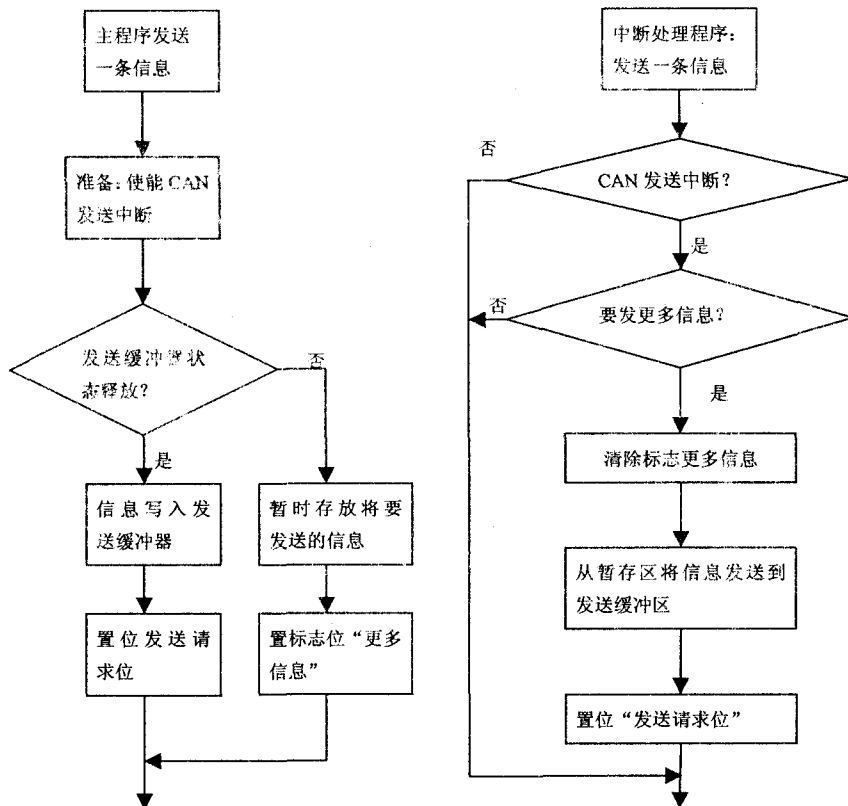


图 2 中断控制流程图:发送一条信息

(6)两台单片机的通信:

两台实验仪分别作为接收端和发送端。

a. 发送端:只发送数据以及接收发送成功的信号,并在 LED 灯上显示 P-NN(NN 为发送成功的数据值)。

b. 接收端:只接受数据以及发送接收成功的信号,并在 LED 灯上显示 0-NN(NN 为接收成功的数据值)。

(7)局域网中对特定单片机的数据采集:

通过设定报文标识符,并发送远程帧向特定的单片机发送数据采集的请求。

对方单片机接收远程帧后,即马上向主机发送数据。

3 系统测试

首先,利用 Keil uVision2 软件进行程序的编辑、调试和编译。

其次,利用 DP-Flash 软件将目标程序下载到实验仪的 Flash 存储器中。

最后,将两台实验仪通过自制的数据线连接起来(将 CAN 总线的接口中的 CAN-L 和 CAN-H 端对应相连),并分别启动两台实验仪。观察实验仪上的 LED 指示灯,验证两台单片机的数据通信。

4 总结

设计并实现了一种基于 CAN 总线的数据采集通信系统。目前 CAN 总线广泛应用于工业控制领域,本数据采集控制系统适用于各种工业控制系统。由于采用 CAN 总线技术,本系统具有功耗低、抗干扰性强和应用简单等特点。

参考文献:

- [1] 广州周立功单片机发展有限公司. CAN 基本知识[EB/OL]. 2001. <http://www.zlgmcu.com>.
- [2] 广州周立功单片机发展有限公司. 控制器局域网(CAN)未来的工业微处理器串行通讯[EB/OL]. 2001. <http://www.zlgmcu.com>.
- [3] 周立功. 增强型 80C51 单片机速成与实战[M]. 北京:北京航空航天大学出版社,2003.
- [4] 饶运涛. 现场总线 CAN 原理与应用技术[M]. 北京:北京航空航天大学出版社,2003.
- [5] Philips Semiconductors. Data Sheet SJA1000 Stand-alone CAN Controller[S]. 2000.
- [6] Philips Semiconductors. Data Sheet TJA1050 High Speed CAN transceiver[S]. 2000.

(上接第 206 页)

文理学院学报:自然科学版,2005,8(4):65-67.

- [2] 李桂影,李 润,任效娥. 专利文献的网上检索[J]. 情报检索,2001(8):67-68.
- [3] Gruber T. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition,1993,5(2):199-220.
- [4] 张宇翔. 知识工程中的本体综述[J]. 计算机工程,2005,31:112-114.
- [5] 杜小勇,李 曼,王大治. 语义 Web 与本体研究综述[J]. 计

算机应用,2004,24(10):14-16.

- [6] 于 红,胡安朋. 对《国际专利分类表》第七版一些变化的研究[J]. 情报理论与实践,2002,25(3):203-206.
- [7] 马海群. 专利信息组织整理中的标准化建设[J]. 情报科学,2000,18(4):293-295.
- [8] 李健蓉. 中国知识产权教程——专利文献与信息[M]. 北京:知识产权出版社,2003.
- [9] 杜小勇,马文峰. 学科领域知识本体建设方法研究[J]. 图书情报工作,2005,49(8):74-78.